Exhibit 1

D582–D585 Nucleic Acids Research, 2004, Vol. 32, Database issue
DOI: 10.1093/nar/gkh089

GELBANK: a database of annotated two-dimensional gel electrophoresis patterns of biological systems with completed genomes

György Babnigg and Carol S. Giometti*

Protein Mapping Group, Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA

Received September 2, 2003; Revised and Accepted October 6, 2003

ABSTRACT

GELBANK is a publicly available database of two-dimensional gel electrophoresis (2DE) gel patterns of proteomes from organisms with known genome information (available at http://gelbank.anl.gov and ftp://bioinformatics.anl.gov/gelbank/). Currently it includes 131 completed, mostly microbial proteomes available from the National Center for Biotechnology Information. A web interface allows the upload of 2D gel patterns and their annotation for registered users. The images are organized by species, tissue type, separation method, sample type and staining method. The database can be queried based on protein or 2DE-pattern attributes. A web interface allows registered users to assign molecular weight and pH gradient profiles to their own 2D gel patterns as well as to link protein identifications to a given spot on the pattern. The website presents all of the submitted 2D gel patterns where the end-user can dynamically display the images or parts of images along with molecular weight, pH profile information and linked protein identification. A collection of images can be selected for the creation of animations from which the user can select sub-regions of interest and unlimited 2D gel patterns for visualization. The website currently presents 233 identifications for 81 gel patterns for *Homo sapiens, Methanococcus jannaschii, Pyrococcus furiosus, Shewanella oneidensis, Escherichia coli* and *Deinococcus radiodurans*.

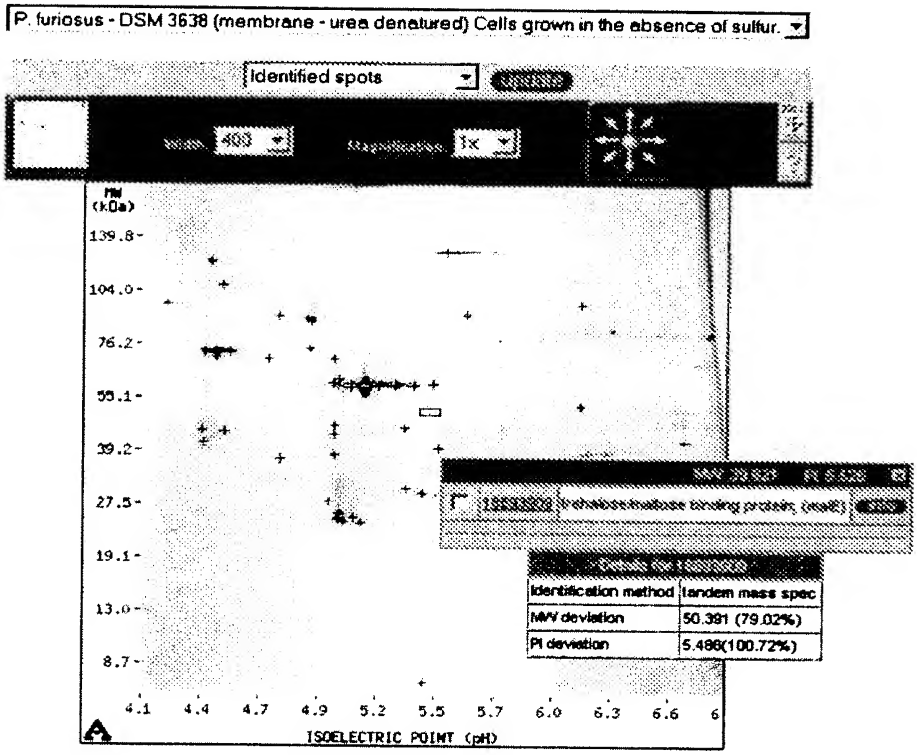
INTRODUCTION

The theoretical proteome of an organism is the collection of potential open reading frames (ORFs). At a given time and in a given state of a cell not all the potential ORFs are expressed. One area of proteomics is the characterization of protein expression profiles under various conditions. Proteome analysis requires the isolation of the complete proteome, separation of complex protein mixtures into discrete protein components, measurement of the relative abundance and identification of each protein component. The most widely used separation method for proteome analysis is currently two-dimensional gel electrophoresis (2DE) (1), where the proteins are separated according to their isoelectric point (pI) in the first dimension and their molecular weight (MW) in the second dimension. Using protein-specific dyes that are stoichiometric with respect to the optical density of the dye and the abundance of the protein, the relative abundance of each protein detected in a 2DE pattern can be determined (2,3). Using calibration proteins, these 2DE gel patterns provide information of relative abundance, MW and pI of isolated proteins in the given pI and MW range. The stained protein spots can be identified by peptide mass fingerprinting by comparing the peptide masses of digested proteins cut out of 2DE gels with the predicted peptide masses from the corresponding genome ORF database.

In the process of analysis of a 2DE gel pattern certain questions arise. Given the complete genome and potential ORF information, what is the most likely identity of a protein at a given location on a gel? What are the predicted proteins resolved in a given pI and MW range? We have calculated theoretical MW and pI values for all the ORFs for 131 organisms with completed genome information. A search engine was built to allow the user to sort proteins based upon their physical characteristics and their annotation data. Theoretical 2DE patterns are also presented with the ability to change both the pI and MW ranges dynamically.

Our laboratory has extensive expertise in 2DE separation of proteins. We have studied protein expression in a variety of organisms including human, rodent and microbial samples. GELBANK not only displays 2DE gel patterns generated in our lab, but it also allows the display and annotation of 2DE gel patterns of registered users with the intent to centralize 2DE gel patterns from around the world. This site is similar to the well-known SWISS-2DPAGE which contains more than 30 reference maps with more than 1000 identifications (4). SWISS-2DPAGE has reference maps for mostly eukaryotes (human, mouse, *Arabidopsis thaliana*, *Dictyostelium discoideum* and *Saccharomyces cerevisiae*), whereas GELBANK has reference maps for mostly prokaryotes (*Methanococcus jannaschii*, *Pyrococcus furiosus*, *Shewanella oneidensis*, *Escherichia coli* and *Deinococcus radiodurans*). The inclusion of Swiss-Prot identification on the ORF display (see below) provides a gateway to the powerful ExPASy server (5).

*To whom correspondence should be addressed. Tel: +1 630 252 3839; Fax: +1 630 252 5517; Email: csgiometti@anl.gov



P. furiosus - DSM 3638 (membrane - urea denatured) Cells grown in the absence of sulfur.

Identified spots

400

1x

Identification method	tandem mass spec
MW deviation	50.391 (79.02%)
PI deviation	5.486(100.72%)

Figure 1. 2DE gel pattern display. The web interface displays a dynamically sizeable image. If profile information is provided, the MW and/or pI values are also displayed dynamically. Spot annotation is displayed by moving the cursor over the spot. Details of the identification (identification method, MW and pI deviation from theoretical values) are available on the same display.

DESCRIPTION

The GELBANK website features six major sections: (i) Proteomes: a collection of interfaces for proteome queries; (ii) 2D Gels: interfaces for 2DE gel pattern queries; (iii) Tools: a collection of tools pertinent to proteome analysis by 2DE; (iv) Upload: web interfaces for registered users for manipulation of their 2DE patterns; (v) Bio-bag: an interface similar to a shopping basket for easy storage of various items from the website; and (vi) FTP site: for the access of data in various formats. The website was tested with Microsoft's Internet Explorer and some features (2DE annotation) work only with this browser.

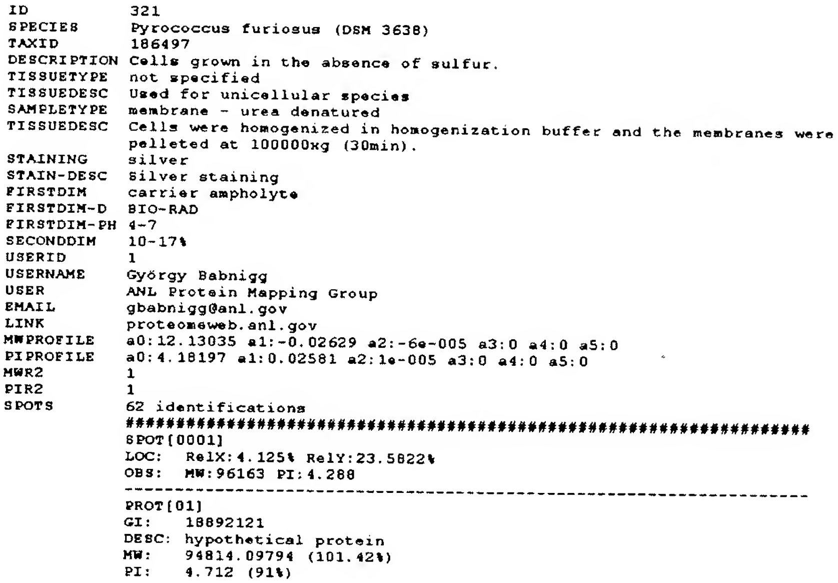
The sections presented above are described in greater detail next.

Proteomes

The potential proteomes of 131 organisms with completed genomes are stored in an ORACLE9i database (currently 464 770 records). The theoretical MW, pI and potential transmembrane regions of every ORF was calculated using the SOSUI server (6). The proteomes can be searched by MW and pI range, ORF annotation, and by sequence fragments or patterns selecting a single organism or a collection of organisms. A single protein or a set of proteins can be selected from the result set and used as input for analysis tools such as multiple alignment (ClustalW) (7), charge distribution at a given pH (see below) or storage for later use in Bio-bag (see below). A summary display for individual entries is also available including links to other publicly available databases (supplementary fig. S1). This display presents a link to the collection of 2DE gel patterns if available for the given organism. If the given ORF is associated with a spot on a given 2DE gel pattern by the annotator, a link is presented to the given pattern(s).

2DE gels

2DE images are uploaded, profiled and annotated by registered users. These high-resolution 2DE images are displayed in a dynamic fashion. 2DE protein patterns are dynamically sizeable from a width of 300 up to 1500 pixels, with image sizes ranging from 30 to 800 kb, respectively. This feature accommodates users with either a slow modem connection or a high-speed connection. The user is able to zoom into any region of the 2DE gel pattern with the corresponding MW and pI scale automatically updated. The gel database can be searched by organism, sample type and annotation. Since gels are profiled in both the MW and pI dimensions the approximation of these properties for a given spot is available (Fig. 1). Spots can be annotated by assigning a single ORF or multiple ORFs from the database. Once a gel has associated annotation, this information is also displayed on the website. If



```
ID            321
SPECIES       Pyrococcus furiosus (DSM 3638)
TAXID         186497
DESCRIPTION Cells grown in the absence of sulfur.
TISSUETYPE  not specified
TISSUEDESC  Used for unicellular species
SAMPLETYPE  membrane - urea denatured
TISSUEDESC  Cells were homogenized in homogenization buffer and the membranes were
            pelleted at 100000xg (30min).
STAINING    silver
STAIN-DESC  Silver staining
FIRSTDIM    carrier ampholyte
FIRSTDIM-D  BIO-RAD
FIRSTDIM-PH 4-7
SECONDDIM   10-17%
USERID      1
USERNAME    György Babnigg
USER        ANL Protein Mapping Group
EMAIL       gbabnigg@anl.gov
LINK        proteomeweb.anl.gov
MWPROFILE   a0:12.13035 a1:-0.02629 a2:-6e-005 a3:0 a4:0 a5:0
PIPROFILE   a0:4.18197 a1:0.02581 a2:1e-005 a3:0 a4:0 a5:0
MWR2        1
PIR2        1
SPOTS       62 identifications
            ####################################################################
            SPOT[0001]
            LOC:  RelX:4.125% RelY:23.5822%
            OBS:  MW:96163 PI:4.288
            --------------------------------------------------------------------
            PROT[01]
            GI:   18892121
            DESC: hypothetical protein
            MW:   94814.09794 (101.42%)
            PI:   4.712 (91%)
```

Figure 2. Example of a database entry. All sequences, 2DE gel patterns and their annotation data can be downloaded from the FTP site (ftp://bioinformatics.anl.gov/gelbank/). An example of an annotated 2DE gel pattern is shown.

the gel has been profiled in both dimensions and an annotation for a given spot is available, the deviation of observed pI and MW of a given spot can be determined by comparing them with the theoretical values. The deviation is displayed in a dynamic fashion along with the identification method used. A link is provided for the detailed ORF information.

GELBANK allows the user to assign multiple identifications for a given spot on a 2DE gel pattern and displays all the associated data. The majority of spots analyzed in our laboratory provide multiple unique hits using tandem mass spectrometry. This might be due to the fact that some of the species are extremophiles and the solubilization method used might not dissociate complexes completely. In other cases the pH gradient used might be too broad for successful separation of proteins with similar physicochemical properties.

The dynamic display of 131 *in silico* patterns is also presented on the website.

Tools

This part of the website has a small collection of proteomics tools relevant to 2DE. In the process of searching the database, a collection of proteins can be selected. These can be saved as a list of sequences (Bio-bag) and can be analyzed later by tools implemented on the website. For example, using ClustalW allows the alignment of these sequences. In certain applications it is important to know the predicted pI and MW of a protein or part of the protein (degradation product). The titration curve and MW of a given complete or partial ORF can be displayed in a dynamic display. A utility allows the simultaneous display of titration curves of an unlimited set of selected ORFs, which is helpful in cases where the optimum pH is searched for the highest difference in charge between two or more proteins (e.g. in ion-exchange chromatography).

Searching GELBANK can also return a list of images. These can also be saved and analyzed by a gel animation tool. Gel animation allows the selection of multiple 2DE patterns, creating a virtual movie. The display of the image collection can be controlled in several ways: (i) change of display time for each frame, (ii) direction of the movie, etc. A web interface allows the user to select a sub-region from each image and play the modified animation. Similar to 2DE patterns, animations can also be saved to Bio-bag (see below).

Some of the tools used were developed at the ProteomeWeb website (8) (http://ProteomeWeb.anl.gov) and shared between the two websites.

Upload

As mentioned above, GELBANK is accepting the upload of 2DE gel patterns from registered users. Users can upload their images using a browser interface. GELBANK currently accepts images only in 8-bit PNG format. The current limitation for file size is 1.4 Mb, which corresponds to a fairly high-resolution image (~1800 × 1800 pixels). An online tool allows the profiling of the uploaded image by selecting markers on the gel and assigning MW and pI values to those points in the image. Several curves are then fitted to the data (first to fifth order polynomials) with the corresponding correlation coefficients and limit values for the given polynomial. The user can select the fit that best describes a given gel, and these data are stored in the ORACLE9i database. A critical feature of any proteome study is the identification of the proteins detected; therefore if annotation data are available, an online tool allows the assignment of an ORF or a collection of ORFs to a given protein spot, thereby linking the 2DE gel pattern to the proteome database. The interface allows the assignment of multiple identifications for a given spot. Assignments can be edited using the same interface. Registered users can add new properties (e.g. identification method for spot ORF assignment, method for separation in the first dimension, etc.) if not listed already, giving great flexibility while maintaining database integrity as all attributes are entered from a list of possibilities. All data are stored in the database for quick retrieval.

Bio-bag

Query interfaces presented on the website allow the storage of objects for later use. These objects currently are: protein sequences, 2DE gel patterns and animations (using intact patterns or sub-regions of patterns). When a registered user is logged into the system these objects are saved in the database and can be recalled later (supplementary fig. S2), otherwise they are deleted after 30 min of inactivity. A few tools are available by selecting a number of these objects: (i) multiple sequence alignment, titration curves for ORFS, (ii) animation for selected 2DE patterns, etc. Contents of the Bio-bag can be deleted if necessary.

FTP site

The entire content of the website is available for download from our FTP site (ftp://bioinformatics.anl.gov/gelbank/). The data are presented in three different formats: (i) simple tab-delimited file (for easy parsing of the information), (ii) flat text format that is easily read by humans as well as by computers, and (iii) ORACLE9i database dumps (Fig. 2). The 2DE patterns are also available for download. The FTP site is updated on weekly basis.

CONCLUSION

A 2DE gel pattern database is presented that ties directly to a proteomics database of species with completed genome information. The public database allows the submission of gel patterns by registered users and their annotation using a web interface. The proteomes can be queried and the resulting data set can be used as an input for some proteomics tools that are pertinent to 2D gel electrophoresis. A new 'shopping basket' is introduced (Bio-bag) that allows the storage of objects presented on the website. Gelbank's content can be downloaded in various forms from the public FTP site. Future releases will include more proteomics tools as well more detailed gel pattern annotations. In addition, once available, Open Source 2DE pattern analysis software will be implemented on the website and users will be able to get analysis reports on submitted gel patterns.

ACKNOWLEDGEMENTS

Online supporting material is available at: http://gelbank.anl.gov/online/NAR2004/. This work was supported by the Office of Biological and Environmental Research through the Microbial Genome Project, US Department of Energy, under Contract No. W-31-109-ENG-38.

REFERENCES

1. O'Farrell,P.Z. and Goodman,H.M. (1976) Resolution of simian virus 40 proteins in whole cell extracts by two-dimensional electrophoresis: heterogeneity of the major capsid protein. *Cell*, **9**, 289–298.
2. Anderson,L.A., Nance,S., Tollaksen,S.L., Giere,F.A. and Anderson,N.G. (1985) Quantitative reproducibility of measurements from Coomassie Blue-stained two-dimensional gels: Analysis of mouse liver protein patters and a comparison of BALB/c and C57 strains. *Electrophoresis*, **6**, 592–599.
3. Giometti,C.S., Gemmell,M.A., Tollaksen,S.L. and Taylor,J. (1991) Quantitation of human leukocyte proteins after silver staining: A study with two-dimensional electrophoresis. *Electrophoresis*, **12**, 536–543.
4. Hoogland,C., Sanches,J.-C., Tonella,L., Binz,P.-A., Bairoch,A., Hochstrasser,D.F. and Appel,R.D. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.*, **28**, 286–288.
5. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
6. Hirokawa,T., Boon-Chieng,S. and Mitaku,S. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.
7. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
8. Babnigg,G. and Giometti,C.S. (2003) ProteomeWeb: A web-based interface for the display and interrogation of proteomes. *Proteomics*, **3**, 584–600.